Identifying imprecise regions for geographic information retrieval using the web

Ross Purves[1*], Paul Clough[2] and Hideo Joho[2]
[1]Department of Geography, University of Zürich, Switzerland
[2]Department of Information Studies, University of Sheffield, UK
*Corresponding author: +41 (0)1 635 6531; rsp@geo.unizh.ch

1. Introduction

Geographic information retrieval is a fast developing area of research. Information retrieval is carried out on unstructured text (commonly web documents) using a structured search request of the form <context><spatial relationship><location>. Thus, a search for "castles in Scotland" searches not only for documents containing the word *castles*, but also for documents which have some *georeference* indicating that they lie within Scotland. This in turn requires that firstly, references to locations must be identified within web documents, and secondly a footprint must be assigned to these references. This process is commonly referred to as *geoparsing* and *geocoding* (Larson, 1996). Such methods work well for well-defined places names which can be mapped directly onto an ontology or gazetteer entry.

Presented with a geographic information retrieval system users are likely to not only enter locations which have well defined footprints but also use common descriptions of location which may not map directly onto entries from an ontology of place. For example, a user may query for "*walks in the Cairngorms"*, where no explicit footprint for the Cairngorms exists. Such a region may be described as an *imprecise region* and user-friendly geographic information retrieval should support the use of such descriptions of location. Vögele et al. (2003) discuss the generation of place name regions (or footprints) for *indeterminate regions* through the use of techniques derived from cognitive geography and spatial metadata.

In this paper an alternative technique for determining imprecise regions is described, based on data mining the web for candidate locations. Boundaries of the imprecise regions are described through the use of density surfaces, although polygon subdivisions may also be derived (Arampatzis et al., 2004). We briefly describe the process of identifying candidate locations and producing density surfaces. Imprecise regions are compared to a set of sketches produced independently to perform a qualitative validation. We then consider how the imprecise regions generated may vary as, for example, a function of weighting functions or query language.

2. Methodology

A number of distinct steps are required to generate imprecise regions – these are described in the following 3 sections.

2.1 Retrieving suitable web pages

We first perform web searches using Google to generate an initial set of 100 documents from which candidate locations can be extracted. To do this, we generate search request (or queries) which are likely to not only retrieve documents about a given place, but also contain several useful geographic references. Search is performed using either *trigger phrases* or searches likely to return pages which are geographically highly populated (e.g. lists of hotels). Trigger phrases attempt to capture geographic relationships and are commonly used in automatic question-answering (Joho and Sanderson, 2000). For example, containment can be represented by a trigger phrase such as *A is a town in B* where B is the imprecise region name, and A is a possible candidate location.

2.2 Geoparsing and geocoding of candidate locations

Having identified a set of web documents geoparsing and geocoding is carried out. Geo-references are extracted using the GATE (General Architecture for TEXT Engineering) information extraction system (Cunningham et al., 2002). Information Extraction (IE) techniques can be used to identify candidate terms for geocoding based on a combination of gazetteer lookup and language-dependent processing. We use two main sources of data for lookup – the SABE (Seamless Administrative Boundaries of Europe) dataset and the Ordnance Survey 1:50,000 Scale Gazetteer. These two datasets contain a total of around 270,000 locations of which about 10% are ambiguous (i.e. not unique entries). When geocoding we can either apply multiple locations to a single reference in these ambiguous cases, or define a *default location* associated with location metadata. Finally, a list of all extracted candidate locations and their frequencies was produced. These frequencies took two forms: so-called *term frequency*, simply a count of how often a geo-reference occurs in all retrieved documents, and *document frequency*, in how many documents a geo-reference occurs.

2.3 Derivation of imprecise regions from point data set

Kernel density surfaces were generated from point data sets, using different weights for the points. Any references to the imprecise region name itself (for example there is a Highland on the south coast of England) were removed. A threshold density value was interactively defined and density surfaces for the imprecise regions generated.

3. Experiments

In a first experiment we asked 5 Swiss geographers to sketch the borders and the heart of the so-called “Mittelland” on a map and compare these sketches with the Mittelland derived from web data. A second set of results illustrate the changing shape of the Highlands as a function of the weighting of the candidate points according to different frequency measures. Finally, we present results for queries on a well defined region (Scotland), where the queries were carried out in English, German and French. Table 1 summarises the information used in querying and the basic properties of the points returned.

Experiment	Query	Restriction on results	Total number of points extracted	Mean number of references to points
Mittelland	Switzerland "the Mittelland"	UK pages only	261	6.1
Highlands	Highlands hotels	UK pages only	474	5.0
Scotland	* is a town in Scotland	UK pages only	672	4.3
Schottland	* ist eine Stadt in Schottland	German pages only	605	4.4
Ecosse	* est une ville en Ecosse	French pages only	353	3.5

Table 1: Queries and number of points extracted in initial search

3.2 First results

Figure 1 shows the 5 sketches produced of the Mittelland and the calculated surface based on the retrieved web documents. Notable features are that the heart of the Mittelland is found in a similar location on 4 of the sketches, though the borders vary somewhat. The density surface captures well the core area of the Mittelland as generated through this simple experiment.

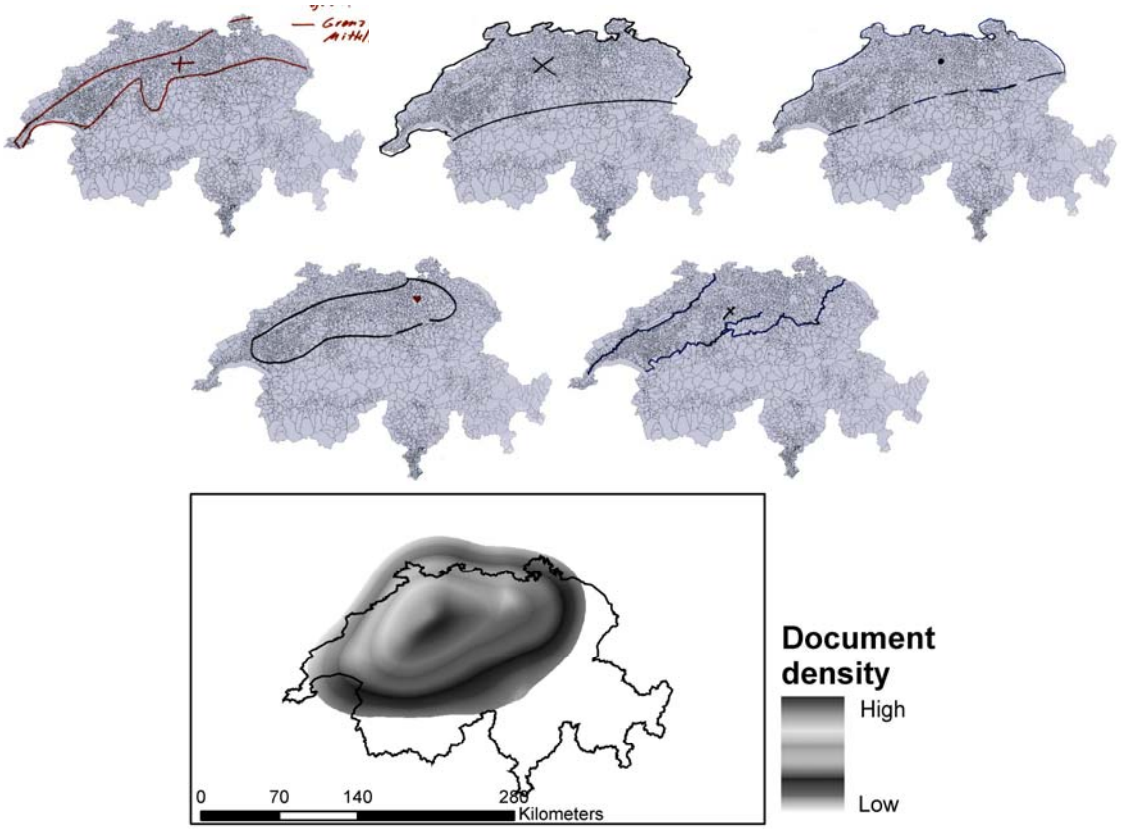
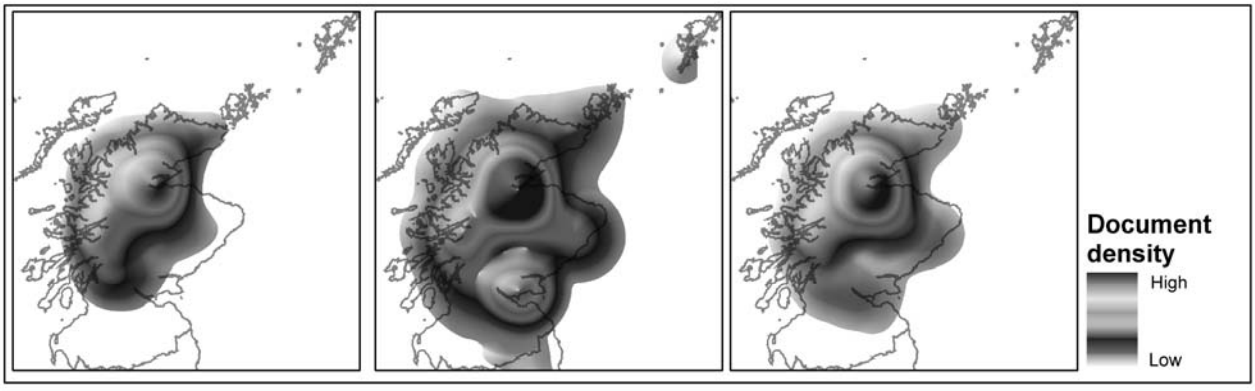


Figure 1: “Mittelland” as defined by 5 subjects and our method. Note crosses indicates “heart of the Mittelland”

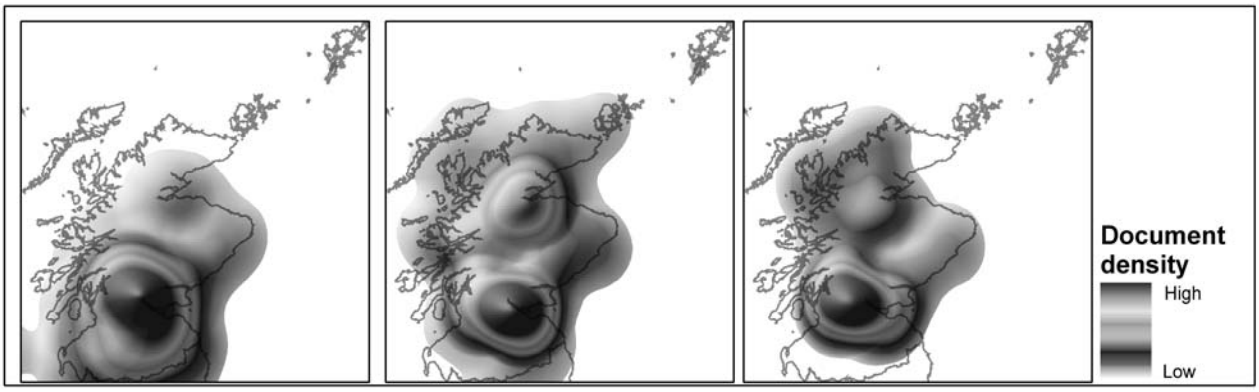
In Figure 2, we attempt to define the area of the Highlands, assuming here that we are interested in an imprecise representation of the Highlands, rather than the unitary local authority. Figure 2 shows the sensitivity of the results to varying the weights of the points used in the density calculation. In Figure 2a, all points are weighted equally and a core area centred on Inverness, and extending north and west is formed. Figure 2b shows how a single false location can bias the resulting surface where points are weighted according to the choice of frequency value. Here, *term frequency* has been used and the results are biased by a few web pages in which a falsely allocated location occurs many times - the significant peak to the south is centred on a village called Cameron in Fife. Cameron is in fact a common Highland surname, and here the geo-referencing has falsely identified it as a location. Finally, in Figure 2c a surface generated with points weighted according to the *document frequency* is shown.



(a) (b) (c)

Figure 2: Sensitivity of the definition of the Highlands to definition of weights in kernel density

In Figure 3 we explore the sensitivity of the definition of an imprecise region to language (and we hypothesise cultural context). Scotland is used as the region and English, German and French queries are generated. The surface generated by the English query (Figure 3a) in fact appears to least well represent the borders of Scotland. We hypothesise that this is because the distribution of web documents referring to Scotland in English tend to reflect the distribution of power and population which, arguably, lies in the central belt. The regions generated by the German and French queries (Figure 3b and c) show much more complete coverage of Scotland, with a clear bias towards the west and north, where many tourist itineraries are to be found.



(a) (b) (c)

Figure 3: Scotland as defined by English, German and French queries respectively

Discussion

In this paper we have demonstrated the application of techniques to data mine locations from the web to generate so-called *imprecise regions*. Initial results, using kernel density surfaces to model the imprecise regions show reasonable agreement with a simple evaluation. The surfaces are strongly dependent on the choice of weighting function, and the language used for the initial trigger query. Further work is also required to explore the use of different interpolation techniques to define imprecise regions. These imprecise regions will be used to define query regions for geographic information retrieval, for example in the generation of egg-yolk models of space for use as query footprints.

Acknowledgements

This research is supported by the EU-IST Project No. IST-2001-35047 (SPIRIT) and the Swiss BBW. OS gazetteer data was provided by Digimap.

References

Arampatzis, A., van Kreveld, M., Reinbacher I., Jones, C.B., Vaid S., Clough, P. Joho, H., Sanderson, M., Benkert, M. and Wolff A. 2004. *Web-Based Delineation of Imprecise Regions*. In Proceedings of Workshop on Geographic Information Retrieval, SIGIR 2004, Sheffield, UK.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

Joho, H., and Sanderson, M. (2000). "Retrieving Descriptive Phrases from Large Amounts of Free Text". In: Proceedings of the 9th International Conference on Information and Knowledge Management, 180-186, McLean, VA: ACM.

Larson, R.R. (1996). Geographic Information Retrieval and Spatial Browsing. In GIS and Libraries: Patrons, Maps and Spatial Information, Linda Smith and Myke Gluck, Eds., University of Illinois.

Vögele, T.; Schlieder, C.; Visser, U. 2003. Intuitive Modelling of Place Name Regions for Spatial Information Retrieval. In: Conference on Spatial Information Theory - COSIT'03, LNCS 2825, Springer, 239-252.

Biography

Ross Purves is a lecturer in the GIS Division of the Department of Geography at the University of Zürich.